

# Examining Socioeconomic Factors Associated with Diabetes Using NHANES Data

Baijia Xu

2026-01-19

## Contents

|                            |   |
|----------------------------|---|
| 1. Introduction . . . . .  | 1 |
| 2. Dataset . . . . .       | 1 |
| 3. Data Cleaning . . . . . | 2 |
| 4. Modeling . . . . .      | 2 |
| 5. Visualization . . . . . | 3 |
| 7. Results . . . . .       | 3 |

## 1. Introduction

Diabetes and socioeconomic disparities are major public health concerns in the United States. Education level is commonly used as a proxy for socioeconomic status, while income-to-poverty ratio is a key indicator for economic stability and health equity. Understanding the relationship between these socioeconomic factors and diabetes prevalence is essential for developing targeted prevention strategies.

**Research Question: How do education and income relate to diabetes prevalence?**

We use data from the National Health and Nutrition Examination Survey (NHANES) to explore the relationship between diabetes status and socioeconomic factors using multinomial logistic regression. The outcome variable is diabetes status (no diabetes, borderline, diabetes), and the predictor variables include income-to-poverty ratio and education level.

## 2. Dataset

NHANES is a nationally representative cross-sectional survey conducted by the Centers for Disease Control and Prevention (CDC). It combines interviews, physical examinations, and laboratory measurements.

For this analysis, we used the following NHANES components:

- 1). Education level and income-to-poverty ratio from Demographics data (`data/DEMO_L.xpt`).
- 2). Diabetes status from Diabetes Questionnaire data (`data/DIQ_L.xtp`).

The analytic sample includes participants with non-missing diabetes, education and income-to-poverty ratio measurements.

### 3. Data Cleaning

Data cleaning and preprocessing were performed to prepare the analytic dataset.

Key steps included:

- 1). Reading XPT format data.
- 2). Merge demographic information with diabetes by sequence number.
- 3). Remove all missing values.
- 4). Drop observations with DMDEDUC3 = 9, which is Unknown.
- 5). Define variable education: labeled as “<9th Grade” if DMDEDUC2=1, “9-11th Grade” if DMDEDUC2=2, “HS Graduate” if DMDEDUC2=3, “Some College” if DMDEDUC2=4, “College+” if DMDEDUC2=5.
- 6). Define variable diabetes\_status: labeled as “Diabetes” if DIQ010=1, “No Diabetes” if DIQ010=2, “Borderline” if DIQ010=3.

### 4. Modeling

We fit a **multinomial logistic regression model** with diabetes as the outcome and education level and income-to-poverty ratio as predictors:

$$\log \left( \frac{P(Y = j)}{P(Y = 0)} \right) = \beta_{0j} + \beta_{1j}X_{\text{edu}} + \beta_{2j}X_{\text{IPR}}$$

where:

$Y$  represents diabetes status with three categories: No Diabetes, Borderline, and Diabetes (reference),

$j$  indexes the diabetes status categories (Diabetes:  $j = 0$ , No Diabetes:  $j = 1$ , Borderline:  $j = 2$ ),

$\beta_{0j}$  is the intercept for category  $j$ ,

$\beta_{1j}$  is the coefficient for education level for category  $j$ ,

$\beta_{2j}$  is the coefficient for income-to-poverty ratio for category  $j$ .

The model fitting results are presented below

```
## Call:  
## multinom(formula = diabetes_status ~ education + INDFMPIR, data = df)  
##  
## Coefficients:  
## (Intercept) education9-11th Grade educationHS Graduate  
## No Diabetes 0.7391356 0.31969278 0.7758026  
## Borderline -1.4043867 -0.01423447 -0.1328210  
## educationSome College educationCollege+ INDFMPIR  
## No Diabetes 0.9298935 1.4913773 0.041414959  
## Borderline -0.0651172 0.4546805 -0.005691361  
##  
## Std. Errors:  
## (Intercept) education9-11th Grade educationHS Graduate  
## No Diabetes 0.1410257 0.1709711 0.1560254  
## Borderline 0.2657182 0.3268562 0.3040241  
## educationSome College educationCollege+ INDFMPIR
```

```

## No Diabetes          0.1547987      0.1696128  0.02597265
## Borderline           0.2994447      0.3207814  0.05294899
##
## Residual Deviance: 6930.36
## AIC: 6954.36

```

## 5. Visualization

Visualizations consist of two parts:

- 1). Stacked bar chart illustrates the distribution of diabetes prevalence across different education levels.
- 2). Density plot compares income-to-poverty ratio distributions among participants with different diabetes status.

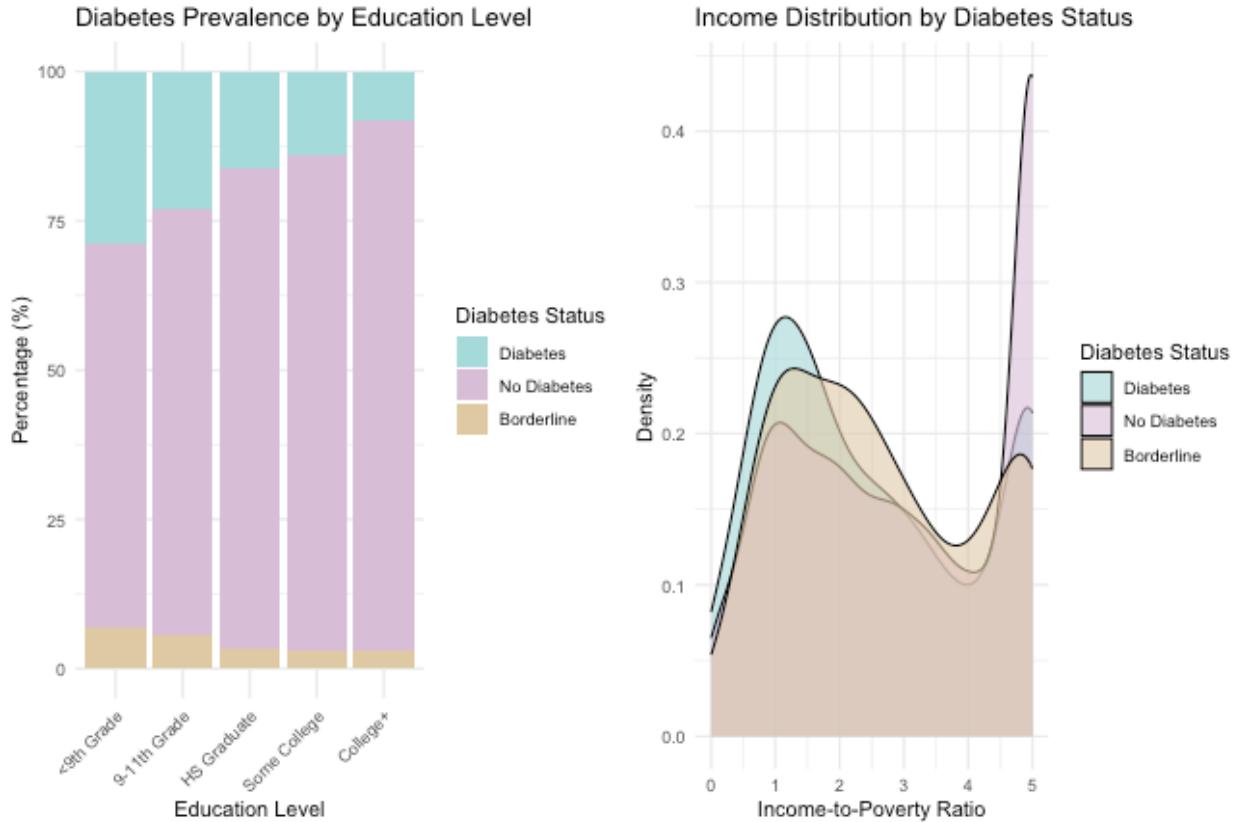


Figure 1 (Left panel) shows the proportion of diabetes status by education level, revealing that higher educational attainment is associated with lower diabetes prevalence. Figure 2 (Right panel) displays the income-to-poverty ratio distributions across diabetes status categories, indicating that individuals without diabetes tend to have higher income-to-poverty ratios compared to those with diabetes or borderline diabetes.

## 7. Results

The multinomial logistic regression results indicate that both education level and income-to-poverty ratio are associated with diabetes status. Higher education levels show positive coefficients for the “No Diabetes” category compared to the reference group, with College+ having the largest coefficient (1.49). For the income-to-poverty ratio, higher values are positively associated with “No Diabetes” (coefficient = 0.041) but negatively associated with “Borderline” diabetes (coefficient = -0.006).